硕士学位论文

基于虚拟物理仿真思考的开放任务求解

PHYSICAL SIMULATION AND REASONING BASED TASK-AGNOSTIC LEARNING

柴士童

哈尔滨工业大学

2020 年 12 月

国内图书分类号：TM301.2　　学校代码：10213
国际图书分类号：62-5　　密级：公开

工程硕士学位论文

基于虚拟物理仿真思考的开放任务求解

硕士研究生：柴士童

导　　师：范晓鹏

申请学位：工程硕士

学　　科：软件工程

所在单位：计算学部

答辩日期：2020 年 12 月

授予学位单位：哈尔滨工业大学

Classified Index: TM301.2
U.D.C: 62-5

Dissertation for the Master's Degree in Technology

PHYSICAL SIMULATION AND REASONING BASED TASK-AGNOSTIC LEARNING

Candidate: CHAI Shitong
Supervisor: Professor FAN Xiaopeng
Academic Degree Applied for: Master of Technology
Specialty: Software Engineering
Affiliation: Faculty of Computing
Date of Defence: December, 2020
Degree-Conferring-Institution: Harbin Institute of Technology

摘　要

本文在Pyrobolearn机器人仿真环境中设计了一个要求机械臂的末端执行器到达指定物体附近的开放任务，并设计了新的强化学习算法用于训练智能体在未获得任务相关的奖励之前在未知环境中学习到可泛化到该开放任务的策略。

在 TD3 算法和 HER 算法的基础上，本文引入了基于局部敏感哈希和计数的奖励和基于正向动力学预测的奖励用于鼓励智能体探索未知环境。

本文提出的斯混合噪声层被用于提供自适应的策略噪声。提出的基于物理仿真引擎仿真时间的奖励被用于在未知任务的环境中鼓励智能体学习有效可泛化的策略，并在获得稀疏奖励的开放任务后快速适应到新的策略。

关键词：元学习；强化学习；机器人学；物理仿真；深度学习

Abstract

A task which requires the end effector of a manipulator to reach a specific body is designed with the help of the robot learning framework Pyrobolearn. New reinforcement learning algorithms are proposed to train an agent to learn in a task-agnostic environment without reward related to the task, where the learnt policy of the agent should be generalizable to the proposed task.

Based on TD3 and HER algorithms, a locality sensitive hashing based counting reward and a forward dynamics prediction model based reward are introduced to encourage the agent to explore in the unknown environment.

A mixed gaussian noise layer is proposed to provide a adaptive policy noise. A reward based on physics simulation time is proposed to encourage the agent to learn a generalizable policy in a task agnostic environment, and adapt to a new policy after given a task setting with sparse reward.

Keywords: Meta Learning, Reinforcement Learning, Robotics, Physics Simulation, Deep Learning

目　录

第 1 章 绪论

1.1 选题背景及研究意义

1.1.1 选题背景

一直以来，强化学习和机器人学都是人工智能研究中的热门领域。在 2008 年，Deepmind 团队基于深度强化学习研发的围棋人工智能系统 AlphaGo Zero 在零知识自我对弈的情况下在几天之内超越了旧的系统 AlphaGo，而 AlphaGo 曾击败了围棋领域中世界公认的专家柯洁等人[?]。这项研究使越来越多的人们开始关注人工智能领域，并使得强化学习成为研究热点。事实上，强化学习已经成为最热门的研究领域之一，并在自动控制、运筹学、机器人学、游戏智能体和无人驾驶等领域中获得了广泛的应用[1]。在这些领域中，机器人学是和前沿强化学习算法关系最密切的领域之一。传统的机器人如机械臂、四足机器人等，可以用强化学习训练得到的智能体进行控制，并在与环境交互过程中根据环境反馈使策略得到进一步的优化。

然而，由于机器人设计和制造的成本较高，通用的多关节机器人通常非常昂贵。而且机器人通常容易在强化学习中的各种随机探索中受到损坏，并导致控制系统和智能体策略在训练过程中出现错误。因此在实际的机器人上进行所有强化学习训练是不现实的[2,3]。为了避免这个问题，可以使用物理仿真引擎对机器人和环境进行建模，并在实际测试智能体之前先在仿真环境中对智能体进行训练。幸运的是，随着强化学习仿真需求的增加，越来越多的针对机器人的仿真环境开始出现，在这些仿真环境中，可以像在真实环境中一样控制机器人的关节、调节各种参数，或获得传感器数据等等，并可以做到在真实环境中难以做到的设定复杂的稀疏奖励、获取碰撞次数、和改变环境的物理参数等操作[4]。

虽然已经有大量的软件系统可以用于在一个环境建模完全精确的情形下解决一个良定义的任务[2]，如何让智能体在面对未知的新环境和未知的新任务后能够有效泛化之前学习到的策略仍然是一个未完全解决的问题。人类可以在陌生的环境中用很少次数的探索自然地掌握大量有效信息，还可以利用已有经验对大量物体进行分类、提高对物理实体运动的预测能力，或创新性地设计工具解决问题。由于人们对大脑的工作原理仍然知之甚少，这个过程通常很难被

数值化为一个单一的奖励函数或强化学习算法。

本课题致力于解决上述问题，即设计算法从而可以训练出能在任务奖励未知的环境中进行探索，获取环境信息，学习基本策略，并在任务确定后快速调节旧策略以适应新任务的智能体。为了实现这个目标，需要利用现有的开源物理仿真引擎、前沿的强化学习算法和具有强大函数拟合能力的深度神经网络来设计新算法。

1.1.2 研究意义

机器人控制对于工业制造有着重要的意义，在工厂流水线上，机械臂常常被设计为只能完成单个简单任务，或需要工人远程控制。虽然它们已经极大地提高了生产效率，减少了对工人们生命财产安全的威胁，但是机械臂由于价格昂贵，仅仅用于单一任务会造成极大的资源浪费。

本课题提出的算法有希望训练出可以在对未知任务奖励的环境进行充分探索之后，快速适应多种不同任务的机器人智能体，从而扩大现有强化学习算法的适用范围，解决更复杂的控制任务，增强机器人智能体泛化策略的能力。

此外，本课题还可以加深对现有强化学习算法在机器人控制中应用价值的理解，可以通过机器人仿真和控制帮助提前发现在应用算法到实际机器人控制时可能出现的问题，可以通过设计和调整深度神经网络进一步了解不同结构的神经网络对智能体性能的影响。

1.2 国内外研究进展

国内外已经有了很多关于让机器人智能体使用工具和泛化已学习的策略到新任务的研究。其中关注工具使用、无监督兴趣导向的探索、模仿学习和自驱动的主动学习与本课题有关。

1.2.1 机械臂工具使用

引导视觉预见[5] 可以使机械臂从人类演示中学习并泛化学习到的能力以在不同的环境中使用工具。这种方法包含动作提案模型和预测模型。其中动作提案模型使用演示动作数据来训练一个自回归的长短时记忆网络模型来根据图像传感器拍摄到的图像数据生成动作。预测模型是基于卷积长短时记忆网络的[6]。预测模型被用于对物理实体的运动进行物理预测，并可以被用于筛选不能完成指定目标的动作序列。训练过程分为基于演示的模仿学习和利用握爪反射的随机自动训练。在测试过程中，指定的目标被定义为像素移动，预测模型

输出的像素位置和真实像素位置的距离被用于评估动作好坏。在指定任务后，动作序列从动作提案模型中采样，并使用交叉熵方法结合预测模型进行优化。实验表明，此种方法可以比单纯模仿学习在新环境中获得更好的泛化性能。

1.2.2 无监督兴趣导向探索

1.2.3 模仿学习

1.2.4 自驱动的主动学习

1.3 研究内容和方法

1.3.1 研究内容

1.3.2 研究方法

第 2 章 相关算法基础

2.1 强化学习基础

2.2 TD3 算法

2.3 HER 算法

2.4 基于局部敏感哈希和计数的探索奖励

2.5 基于正向动力学预测的探索奖励

第 3 章 实验环境

Gym 和 Pyrobolearn[?]

3.1 Gym

?

3.2 Pyrobolearn

?

3.3 环境参数

?

3.4 系统设计

系统由策略对象和奖励函数构成

第 4 章 混合高斯噪声层

4.1 网络结构

4.2 实验结果

第 5 章 仿真时间奖励

5.1 工作原理

5.2 实验结果

结　论

本文结合了 TD3 算法、HER 算法和基于 LSH 和计数的奖励、基于正向动力学预测的奖励，并提出了混合高斯噪声层和将物理仿真时间作为于未知任务的探索奖励。本文中的算法可以很好地用于训练可适应到给定开放任务的智能体，表现出与现有探索奖励相当的性能。

参考文献

[1] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An Open Urban Driving Simulator[J]. Conference on Robot Learning, 2017 : 1-16.

[2] Toussaint M, Allen K R, Smith K A, et al. Differentiable Physics and Stable Modes for Tool-Use and Manipulation Planning.[C] // Robotics: Science and Systems XIV : Vol 14. 2018.

[3] Todorov E, Erez T, Tassa Y. MuJoCo: A physics engine for model-based control[C] // 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012 : 5026-5033.

[4] Savva M, Malik J, Parikh D, et al. Habitat: A Platform for Embodied AI Research[C] // 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019 : 9339-9347.

[5] Xie A, Ebert F, Levine S, et al. Improvisation through Physical Understanding: Using Novel Objects As Tools with Visual Foresight[C] // Robotics: Science and Systems XV : Vol 15. 2019.

[6] Shi X, Chen Z, Wang H, et al. Convolutional LSTM Network: a machine learning approach for precipitation nowcasting[C] // NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 : Vol 28. 2015 : 802-810.

附录 A 算法代码

A.1 训练代码

A.2 测试代码

索引

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于虚拟物理仿真思考的开放任务求解》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：　　日期：　　年　月　日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

致　谢

衷心感谢导师 范晓鹏 教授对本人的精心指导。他的言传身教将使我终生受益。

感谢哈工大 LaTeX 论文模板 hiThesis 。